2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

The Filtering Effect of Face Masks in their Detection from Speech

Adria Mallol-Ragolta[1*], Shuo Liu[1], and Björn W. Schuller[1,2]

***Abstract*— Face masks alter the speakers' voice, as their intrinsic properties provide them with acoustic absorption capabilities. Hence, face masks act as filters to the human voice. This work focuses on the automatic detection of face masks from speech signals, emphasising on a previous work claiming that face masks attenuate frequencies above 1 kHz. We compare a paralinguistics-based and a spectrograms-based approach for the task at hand. While the former extracts paralinguistic features from filtered versions of the original speech samples, the latter exploits the spectrogram representations of the speech samples containing specific ranges of frequencies. The machine learning techniques investigated for the paralinguistics-based approach include Support Vector Machines (SVM), and a Multi-Layer Perceptron (MLP). For the spectrograms-based approach, we use a Convolutional Neural Network (CNN). Our experiments are conducted on the Mask Augsburg Speech Corpus (MASC), released for the Interspeech 2020 Computational Paralinguistics Challenge (COMPARE). The best performances on the test set from the paralinguistic analysis are obtained using the high-pass filtered versions of the original speech samples. Nonetheless, the highest Unweighted Average Recall (UAR) on the test set is obtained when exploiting the spectrograms with frequency content below 1 kHz.**

I. INTRODUCTION

Respiratory viruses, such as the *Coronavirus Disease 2019* (COVID-19), are transmitted via direct contact with an infected person or a contaminated surface, and through respiratory droplets containing the virus, which can be suspended in the air for a long time and over a long distance [1]. According to the *World Health Organisation* (WHO), current evidences support the hypothesis that COVID-19 mainly spreads by respiratory droplets among people in close contact when coughing, sneezing, speaking, singing or breathing heavily. As a consequence, Governments worldwide have ruled on the obligatoriness of wearing face masks in public transports, public spaces, frequented streets, shops, and even in workplaces to control the spread of COVID-19.

The need to check the compliance with this precaution measure has motivated the use of new digital solutions for the automatic detection of face masks from both visual and acoustic signals. Analysing the information embedded in the visual modality, hybrid models combining deep and classical machine learning techniques [2], and transfer learning-based approaches [3], [4] have been proposed. Focusing on the acoustic modality, the speech changes produced by wearing a mask [5], the use of deep convolutional neural networks from the spectrograms of the audio signals [6], and solutions based on transfer learning [7], [8] have also been investigated.

The materials used to produce face masks give them specific properties in terms of thickness and porosity. These properties cause the absorption of frequency content [9] when placing the mask in front of a sound source. This phenomenon explains the changes produced in our voice when wearing a face mask, as it behaves as a filter. A recent study claims that face masks attenuate frequencies above 1 kHz [10]. In the problem of face mask detection from speech, our hypothesis is that the exploitation of the frequency content above this cut-off frequency should be more discriminative. For this, we compare the performance of face mask detection models trained with the paralinguistic information extracted from filtered versions of the original speech signals. Furthermore, as the filtering effect can also be observed in the frequency domain, we investigate the performance of models based on a *Convolutional Neural Network* (CNN) to exploit the spectrograms of the original audio signals.

The rest of the paper is laid out as follows: Section II describes the dataset analysed in this work, while Section III details the methodology followed. Section IV summarises and interprets the performance of the models trained, and Section V concludes the paper and suggests potential directions for future works.

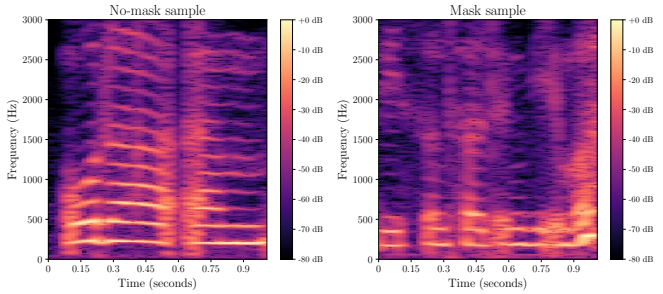
II. DATASET

The *Mask Augsburg Speech Corpus* (MASC), released for the Mask Sub-Challenge of the Interspeech 2020 *Computational Paralinguistics Challenge* (COMPARE) [11], is explored in this work. This corpus contains speech samples from 32 German native speakers (16 f, 16 m) performing different tasks with and without wearing the Sentinex Lite surgical mask from Lohmann and Rauscher. The tasks proposed included answering some questions, reading specific words, describing pictures, and drawing a picture and talking about it. The total duration of the dataset is 10 h 9 min 14 sec. The audio files released were sampled at 16 kHz, and segmented into frames of 1 second length without overlap. The total number of samples available per class and partition are summarised in Table I.

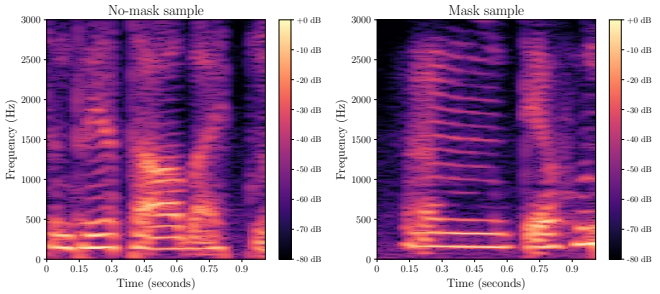
The main hypothesis of this work is based on the filtering effect to the human voice caused by face masks. As a preliminary investigation into this corpus, we computed the spectrograms of different audio frames to visually analyse whether face masks absorb the high frequencies in the speech samples available. A selection of the spectrograms computed

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826506 (sustAGE).

[1] Adria Mallol-Ragolta, Shuo Liu, and Björn W. Schuller are with the Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Augsburg, Germany.
[2] Björn W. Schuller is also with the Group on Language, Audio, & Music, Imperial College London, London, UK.
* Corresponding author: adria.mallol-ragolta@uni-a.de



(a) Spectrograms computed from a female participant while speaking without (left) and with (right) a surgical face mask.



(b) Spectrograms computed from a male participant while speaking without (left) and with (right) a surgical face mask.

Fig. 1: Visualisation of the spectrograms extracted from the original audio samples available in the *Mask Augsburg Speech Corpus* (MASC) from a female (Figure 1a) and a male (Figure 1b) participant.

TABLE I: Summary of the total number of audio samples available in the *Mask Augsburg Speech Corpus* (MASC) per class and partition.

	Train	Devel	Test	$\sum$
No-mask	5 353	6 666	5 553	17 572
Mask	5 542	7 981	5 459	18 982
$\sum$	10 895	14 647	11 012	36 554

is depicted in Figure 1. The spectrograms computed belong to a female (cf. Figure 1a) and a male (cf. Figure 1b) participant with and without wearing a surgical face mask while speaking. Comparing the spectrograms from each participant, less energy in the frequencies about 1 kHz can be observed in the spectrograms corresponding to the samples recorded while wearing a face mask. Hence, the cut-off frequency of 1 kHz seems a suitable threshold for our experiments.

III. METHODOLOGY

This section describes the methodology followed in this work. Sections III-A and III-B detail the approaches followed when analysing the paralinguistic information extracted, and the spectrograms computed from the audio signals, respectively. Section III-C synthesises the parameters and procedures used to train our neural network-based models.

A. Paralinguistics-based Approach

To investigate the filtering effect from a paralinguistic perspective, we first create filtered versions of the original audio signals. We opt for a Butterworth filter [12], because of its flat magnitude of the frequency response in the pass-band. For the purposes of this study, we design a low-pass and a high-pass filter with a cut-off frequency at 1 kHz. Different order Butterworth filters are considered during the design phase, and their magnitude of the frequency response is depicted in Figure 2. Because of their quicker roll off and their shorter transition between the pass-band and the stop-band, we select the low-pass and the high-pass 20^{th}-order Butterworth filters.

The next step is the extraction of the paralinguistic features from the original, low-pass, and high-pass versions of the audio samples. We extract the 6 373 features available from the ComParE feature set [13], [14] using the 3.0 public

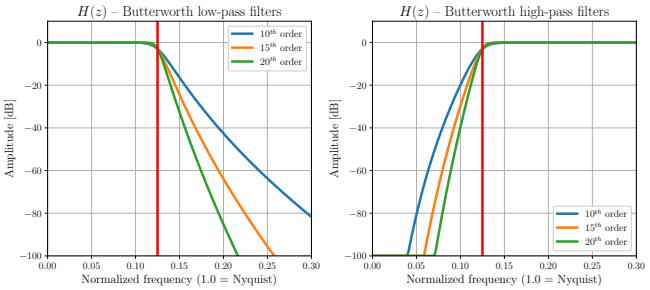


Fig. 2: Magnitude of the frequency response of the low-pass (left) and high-pass (right) 10^{th}-, 15^{th}- and 20^{th}-order Butterworth filters with a cut-off frequency at 1 kHz.

release of openSMILE [15], [16]. We apply standardisation to the training features, so they are zero-mean and unit-variance. To reduce the dimensionality of the features while preserving the information that contributes the most, we apply dimensionality reduction with *Principal Component Analysis* (PCA), keeping 90 % of the variance. The number of components kept depends on the version of the audio sample used. Both the standardisation and the PCA parameters are only computed on the training features, and stored, so these can be applied to the testing features off-line.

We compare two different techniques to model the paralinguistic features: the first one uses a *Support Vector Machines* (SVM), and the second one, a *Multi-Layer Perceptron* (MLP). The SVM uses a linear kernel with a regularisation parameter $C \in [10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}]$, optimised on the development partition of the dataset. The MLP implements three-stacked *Fully Connected* (FC) layers with 64, 16, and 1 neurons, respectively. While the first two FC layers use the *Rectified Linear Unit* (ReLU) as the activation function, the last layer uses the Sigmoid function as the activation, so the output of the network belongs to the range $[0, 1]$. If the output is greater than .5, we consider the input sample to be recorded while wearing a face mask.

B. Spectrograms-based Approach

The starting point of this approach is the computation of the *Short-Time Fourier Transform* (STFT) of the original audio samples. We compute the STFT using the *librosa* library [17], with a window length and a hope size of 2 048 and 256 samples, respectively. The magnitude of the STFT

TABLE II: Summary of the UAR scores (in percentage) computed from the SVM and the MLP models trained following the paralinguistics-based approach. The 6 373 features of the ComParE feature set are extracted from the original audio samples, and their low-pass and high-pass filtered versions, which are obtained using a low-pass and high-pass 20^{th}-order Butterworth filter, respectively. The extracted features are standardised, and dimensionality reduction is applied using PCA, keeping 90 % of the variance.

Models	**SVM-based**		**MLP-based**	
Partition	Devel	Test	Devel	Test
Original signals	62.67	67.03	62.87	67.13
LP signals	58.45	61.70	58.23	61.40
HP signals	63.13	**67.22**	62.49	**67.74**

is computed and normalised, so it ranges between 0 and 1 before generating the spectrograms. In our experiments, we treat the spectrogram representations of the original audio files as images, so these can be exploited using a CNN. In this regard, the frequency scale selected to display the spectrograms modifies their visual appearance. Thus, we aim to compare the performance of networks exploiting spectrogram representations using a linear and a logarithmic frequency scale. Furthermore, to investigate the filtering effect in the spectrograms, we focus our analysis on specific frequency ranges. Our experiments compare the use of spectrograms with frequency content over all the range available (i. e. , from 0 kHz to 8 kHz), with spectrograms containing only the frequencies below 1 kHz (i. e. , from 0 kHz to 1 kHz), and above 1 kHz (i. e. , from 1 kHz to 8 kHz). The spectrograms generated are saved as images of 256×256 pixels for further processing.

The CNN-based network implemented to exploit the information available from the spectrograms of the audio samples has two main blocks. The first block extracts deep-learnt feature representations of the input spectrograms, while the second block is responsible for the actual classification. The first part of the network is composed of two convolutional blocks with 6 and 16 channels, respectively, with a square kernel of 5×5, and a stride of 2. The output of both convolutions is forwarded to a ReLU activation function, and a 2×2 max-pooling layer. Analogously to the MLP implemented in Section III-A, the classification block of our network has three-stacked FC layers with 64, 16, and 1 neurons, respectively. The first two FC layers use ReLU as the activation function, while the last FC layer uses a Sigmoid function as the activation. This way, the output of the network belongs to the range $[0, 1]$. If the output is greater than .5, we consider the input sample to be recorded while wearing a face mask.

C. Networks Training

The MLP and the CNN-based network described in Sections III-A and III-B, respectively, are trained under the exact same conditions. Both networks use the *Mean Squared Error* (MSE) as the loss to optimise, and Adam as the optimiser, which implements a fixed learning rate of 10^{-3}. The network parameters are updated every 256 samples, and trained during a maximum of 100 epochs. An early stopping mechanism is also implemented to stop training when the validation loss does not improve for 10 consecutive epochs. This allows determining the number of epochs required for the network to be trained while preventing overfitting. Consequently, this parameter defines the amount of training required for the networks at the testing stage, when considering samples from both the training and development partitions as training material.

TABLE III: Summary of the UAR scores (in percentage) computed from the CNN-based networks trained with the spectrograms generated using a linear or a logarithmic frequency scale following the spectrograms-based approach. The scenario in which the input spectrograms contain frequency content over all the range available (i. e. , from 0 kHz to 8 kHz) is referred as *Original spectrograms* in the table. The scenarios in which the input spectrograms only contain frequency content below 1 kHz (i. e. , from 0 kHz to 1 kHz), and above 1 kHz (i. e. , from 1 kHz to 8 kHz) are referred as *LF spectrograms* and *HF spectrograms*, respectively.

Frequency Scale	**Linear**		**Logarithmic**	
Partition	Devel	Test	Devel	Test
Original spectrograms	60.49	58.28	67.05	69.15
LF spectrograms	60.76	**66.42**	65.83	**70.70**
HF spectrograms	60.18	59.22	60.02	57.99

IV. Experimental Results

For a fair comparison with related works in the literature, we assess the performance of our models by computing the *Unweighted Average Recall* (UAR) between the labels inferred and the ground truth. The results obtained following the paralinguistics-based approach (cf. Section III-A) and the spectrograms-based approach (cf. Section III-B) are synthesised in Tables II and III, respectively. The performance of our models is comparable with the model performances reported in the baseline paper of the ComParE 2020 Challenge [11], although their best model on the test set scored a UAR of 71.8 % using a fusion of best approach.

Analysing the results obtained with the paralinguistics-based approach (cf. Table II), we can state that the performance of both SVM- and MLP-based models is similar. Regardless of the technique, we observe that the best UAR scores are obtained when exploiting the paralinguistic features extracted from the high-pass filtered versions of the original speech signals, while the worst UAR scores, from the low-pass filtered versions. The former scored a UAR of 67.22 % and 67.74 %, while the latter achieved a UAR of 61.70 % and 61.40 % with the SVM- and MLP-based models, respectively. These results support the filtering effect of face masks, as the speech signals with frequency content above 1 kHz contain more discriminative information for the task at hand. The results obtained show a small performance difference when using the original speech signals, and their high-pass

filtered versions. We hypothesise this can be attributed to the combination of applying PCA on the paralinguistic features, preserving the features with the highest variance only, and the capabilities of the model, which might have learnt to downsize the effect of the low frequency-related features.

When exploiting the salient information from the spectrogram representations of the speech signals (cf. Table III), we observe that the best performances are obtained with the spectrograms containing only the frequencies below 1 kHz, scoring a UAR of 66.42 % and 70.70 % when using the linear and the logarithmic frequency scales, respectively. As the fundamental frequency of average male and female speakers lies below 500 Hz, the highest energies in the spectrogram can be expected at the low frequencies. Consequently, spectrograms with frequency content below 1 kHz might contain richer information to be extracted by the network. Furthermore, this result can be interpreted as follows: the filtering effect of face masks impacts the energy of the spectrogram representations of the speech signals, in such a way that is detectable by CNNs. In line with these observations, our experiments support the use of spectrograms with the logarithmic frequency scale, in which low-frequency bins cover a narrow range of frequencies, while the high-frequency bins, a wider range.

V. CONCLUSIONS

This work investigated the automatic detection of face masks from speech signals, emphasising on the filtering effect caused by face masks to the frequencies above 1 kHz. The MASC dataset released for the Interspeech ComParE 2020 edition was used to test our hypothesis, and to assess the performance of our models. When focusing the analysis on the use of paralinguistic features, the results obtained indicated that high-pass filtered versions of the speech signals contained more salient information for the task at hand. This supported our hypothesis, as the features of the speech samples containing only frequencies above 1 kHz were more discriminative. Nevertheless, when exploring the use of the spectrogram representations of the speech signals, spectrograms containing frequencies below 1 kHz achieved the highest UAR scores. This result suggested that face masks impacted the energy of the spectrogram representations of the speech signals.

One of the limitations of the dataset used in this work is that the speech samples were only recorded using surgical face masks. The different types of face masks are made of different materials, and, therefore, have different absorption properties. As future work, the collection of a similar dataset using FFP2 masks could be considered with the aim to assess how the filtering effect of this type of masks impacts their automatic detection from speech. Further research could explore transfer learning techniques in this problem, using state-of-the-art pre-trained CNNs to extract deep-learnt representations of the spectrograms. Work in this direction could help to assess the benefits of using pre-trained or specific CNNs when training face mask detection models from speech signals, and to determine the optimal approach.

REFERENCES

[1] Editorial, "COVID-19 transmission – up in the air," *The Lancet. Respiratory Medicine*, 2020.

[2] Mohamed Loey, Gunasekaran Manogaran, Mohamed Hamed N Taha, and Nour Eldeen M Khalifa, "A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic," *Measurement*, vol. 167, pp. 108288, 2021.

[3] G. Jignesh Chowdary, Narinder Singh Punn, Sanjay Kumar Sonbhadra, and Sonali Agarwal, "Face Mask Detection Using Transfer Learning of InceptionV3," in *Proc. of the International Conference on Big Data Analytics*, Sonepat, India, 2020, pp. 81–90, Springer.

[4] Mohamed Loey, Gunasekaran Manogaran, Mohamed Hamed N Taha, and Nour Eldeen M Khalifa, "Fighting against covid-19: A novel deep learning model based on yolo-v2 with resnet-50 for medical face mask detection," *Sustainable Cities and Society*, vol. 65, pp. 102600, 2021.

[5] Claude Montacié and Marie-José Caraty, "Phonetic, Frame Clustering and Intelligibility Analyses for the INTERSPEECH 2020 ComParE Challenge," in *Proc. of Interspeech*, Shanghai, China, 2020, pp. 2062–2066, ISCA.

[6] Jeno Szep and Salim Hariri, "Paralinguistic Classification of Mask Wearing by Image Classifiers and Fusion," in *Proc. of Interspeech*, Shanghai, China, 2020, pp. 2087–2091, ISCA.

[7] Tomoya Koike, Kun Qian, Björn W. Schuller, and Yoshiharu Yamamoto, "Learning Higher Representations from Pre-Trained Deep Models with Data Augmentation for the COMPARE 2020 Challenge Mask Task," in *Proc. of Interspeech*, Shanghai, China, 2020, pp. 2047–2051, ISCA.

[8] Maxim Markitantov, Denis Dresvyanskiy, Danila Mamontov, Heysem Kaya, Wolfgang Minker, and Alexey Karpov, "Ensembling End-to-End Deep Models for Computational Paralinguistics Tasks: ComParE 2020 Mask and Breathing Sub-Challenges," in *Proc. of Interspeech*, Shanghai, China, 2020, pp. 2072–2076, ISCA.

[9] F Alton Everest and Ken C Pohlmann, *Master Handbook of Acoustics*, McGraw-Hill Education, 2015.

[10] Ryan M Corey, Uriah Jones, and Andrew C Singer, "Acoustic effects of medical, cloth, and transparent face masks on speech signals," *The Journal of the Acoustical Society of America*, vol. 148, no. 4, pp. 2371–2375, 2020.

[11] Björn W. Schuller, Anton Batliner, Christian Bergler, Eva-Maria Messner, Antonia Hamilton, Shahin Amiriparian, Alice Baird, Georgios Rizos, Maximilian Schmitt, Lukas Stappen, Harald Baumeister, Alexis Deighton MacIntyre, and Simone Hantke, "The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing & Masks," in *Proc. of Interspeech*, Shanghai, China, 2020, pp. 2042–2046, ISCA.

[12] Stephen Butterworth, "On the Theory of Filter Amplifiers," *Experimental Wireless and the Wireless Engineer*, vol. 7, pp. 536–541, 1930.

[13] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, Marcello Mortillaro, Hugues Salamin, Anna Polychroniou, Fabio Valente, and Samuel Kim, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proc. of Interspeech*, Lyon, France, 2013, pp. 148–152, ISCA.

[14] Björn Schuller, Stefan Steidl, Anton Batliner, Julia Hirschberg, Judee K. Burgoon, Alice Baird, Aaron Elkins, Yue Zhang, Eduardo Coutinho, and Keelan Evanini, "The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language," in *Proc. of Interspeech*, San Francisco, CA, USA, 2016, pp. 2001–2005, ISCA.

[15] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – The Munich Versatile and Fast Open-source Audio Feature Extractor," in *Proc. of the* 18^{th} *International Conference on Multimedia*, Firenze, Italy, 2010, pp. 1459–1462, ACM.

[16] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in *Proc. of the* 21^{st} *International Conference on Multimedia*, Barcelona, Spain, 2013, pp. 835–838, ACM.

[17] Brian McFee, Matt McVicar, Stefan Balke, Carl Thomé, Colin Raffel, Dana Lee, Oriol Nieto, Eric Battenberg, Dan Ellis, Ryuichi Yamamoto, Josh Moore, Rachel Bittner, Keunwoo Choi, Pius Friesch, Fabian-Robert Stöter, Vincent Lostanlen, Siddhartha Kumar, Simon Waloschek, Seth Kranzler, Rimvydas Naktinis, Douglas Repetto, Curtis "Fjord" Hawthorne, CJ Carr, Waldir Pimenta, Petr Viktorin, Paul Brossier, João Felipe Santos, Jackie Wu, Erik Peterson, and Adrian Holovaty, "librosa/librosa: 0.8.0," 2020.